面向智慧知识服务的科技文献大数据体系建设*

吴振新[1,2] 钱力[1,2] 谢靖[1,2] 常志军[1,2] 许丽媛[1] 赵艳[1,2]
[1] 中国科学院文献情报中心 北京 100190 [2] 中国科学院大学经济与管理学院图书情报与档案管理系 北京 100190

摘 要:[目的/意义]探索构建文献情报大数据知识资源体系,支撑面向多领域的智慧知识服务。[方法/过程]基于 AI 应用需求,借鉴业界经验,梳理现有资源体系的问题,从多层次多维度扩展资源体系;构建可靠数据处理流程和计算平台,支持高效数据采集和处理;研发智能化数据治理工具,实现知识资源的有效治理,确保提供高质量数据资源。[结果/结论]已初步形成覆盖多类型、多学科的科技文献大数据知识资源体系,构建完成高度自动化的数据采集治理流程,实施多重数据质量控制,积累数亿高质量数据,且为多个知识服务提供数据支撑。

关键词:科技大数据 知识资源体系 数据汇聚 智慧知识服务
分类号:G250.7
DOI:10.13266/j.issn.0252-3116.2020.24.008

1 前言

人工智能(AI)和大数据已经成为影响社会各个领域的通用技术,正在颠覆和改变它所触及的每一个行业。同样,它们也以一种全新的模式推动了科学研究的突破[1],并为知识服务提供了一种全新范式,从而激发出智慧知识服务的强烈需求。

智慧知识服务,即充分利用 AI + 大数据技术搭建智能文献情报系统,让科技情报工作成为灵活运转的以智能文献情报系统为核心的"数据清洗厂""信息加工厂""知识生成厂"与"决策制定厂",使科技情报工作能够快速洞悉变化、凝练问题、聚焦目标、形成解决方案,极大地弥补人类智能上的不足,增强人们应对复杂问题与任务的能力。

搭建智能文献情报系统,对原有文献情报数据体系提出了新的需求和挑战,我们需要面向 AI 应用需求重新梳理原有的数据体系以支持这种新技术的应用,并为最终的智慧知识服务提供知识型数据支撑。中国科学院文献情报中心(简称 NSLC)面向未来发展提出了"建设 AI + 智慧知识服务生态体系"的目标,作为智慧知识服务生态的有机组成,科技文献大数据体系建设将成为打造未来核心竞争力的重要内容。

2 科技文献大数据体系设计

科技文献大数据体系主要包括数据体系、管理平台以及围绕两者的标准规范及技术方法。基于需求驱动设计的理念,笔者首先分析支持智慧知识服务的数据需求,并据此形成设计思路,完成体系框架的设计。

2.1 支撑智慧知识服务的数据需求分析

2.1.1 AI 应用需求

智慧知识服务是以 AI 应用为特点的,我们需要分析 AI 应用对于大数据体系的影响和需求。

数据科学家 R. Monica 针对 AI 应用提出了 AI 需求层次论[2],AI 应用的流程从底层的数据采集、存储、清洗到逐步应用 AI,每一阶段都对应着不同的数据和处理需求,整个流程难度逐步递进(见图 1)。她认为:扎实的数据基础是第一要素,可靠的数据流程、便捷的数据工具也是 AI 应用的关键。

计算机领域普遍认为 AI 应用的 3 个因素是算法、算力、数据,其中数据是核心竞争力。想从 AI 中获益,需要大量的训练数据[3]。

艾瑞咨询在近期发布的《2020 中国 AI 基础数据

* 本文受"中国科学院文献情报中心成立七十周年主题论坛与纪念文集出版"项目资助出版。
作者简介:吴振新(ORCID:0000-0003-4966-1961),研究馆员,博士生导师,;钱力(ORCID:0000-0002-0931-2882),研究馆员,硕士生导师;谢靖(ORCID:0000-0001-6698-1786),副研究馆员,硕士生导师;常志军(ORCID:0000-0001-9211-8599),副研究馆员,硕士生导师;许丽媛(ORCID:0000-0002-8326-4372),馆员,通讯作者,E-mail:xuly@mail.las.ac.cn;赵艳(ORCID:0000-0002-0515-1954),研究馆员,博士,硕士生导师。
收稿日期:2020-11-05 修回日期:2020-12-20 本文起止页码:63-72 本文责任编辑:徐健

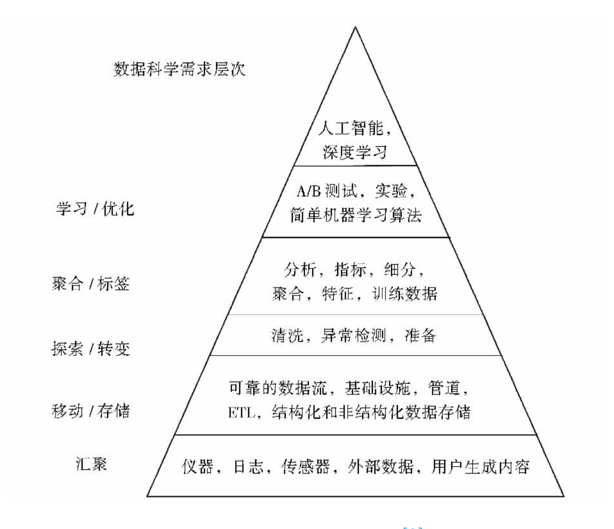


图 1　AI 需求层次论[2]

服务行业发展报告》中指出：目前人工智能商业化在算力、算法和技术方面基本达到阶段性成熟，想要更加落地，解决行业具体痛点，需要大量经过标注处理的相关数据做算法和模型的训练支撑[4]。

由此可见，高质量数据（特别是标注数据）、可靠的数据治理流程、多样化治理工具是 AI 应用的关键。

2.1.2　分析和借鉴国际出版商的智慧知识服务

大数据与人工智能技术的应用，也推动着科技知识服务模式的改变[5,6]，国际出版社借助数据优势，率先利用 AI 技术探索新型知识服务。笔者调研了他们所开展的智慧知识服务作为我们进行体系架构设计时的参考和借鉴。

Elsevier[7] 构建了涵盖数据、证据、工具与智慧服务的新型科研生态，并发布了一系列数字化、知识化工具；Digital Science[8] 则面向科研全流程，提出了一种全新的科研信息服务模式，形成了包括研究人员、科研机构、基金项目与出版物这 4 个维度的数据体系，并研发了多种智能工具；Taylor & Francis，除了自有数据，还集成了其他多种来源数据，其发布的知识图谱工具 Wizdom. ai[9] 中涵盖出版物、专利、作者、机构、概念、事实等数亿数据。

可以看出他们的工作也是主要集中在数据、平台、工具这 3 个方面。

2.1.3　面向问题的科技大数据体系扩展需求分析

基于上述调研分析，笔者仔细审视了原有体系架构，要应用 AI 技术和支持智慧知识服务，笔者需要提供更为丰富的高质量数据（包括标注数据集）、可靠的数据治理流程、多样化治理工具，同时要解决多来源数据融汇、数据标引、深度知识融合、领域知识图谱构建等难题，而这些需要原有数据体系从多个层面进行扩展调整。

从数据层面，要丰富和扩展现有的科技文献数据体系，形成多层次的面向不同功能的数据群。因此，不但要包括传统的科技文献基础数据群，还要建立数据治理支撑数据群，用以支持智能化数据加工并实现智能化数据增值，同时要建立科技知识关联计算数据群，用以支撑知识计算实现智能化的知识生成和决策制定。

从流程层面，改造原有的数据处理流程，以数据治理为核心，嵌入 AI 及大数据技术，提升整个数据流程的可靠性和高效性，既要保障大数据高效采集，又要确保高效计算，使得数据得到及时治理和更新，能够通过智能化数据治理获得支撑智慧知识服务所需的高质量数据。

从多样化智能工具层面，引入 AI + 知识挖掘等新技术新方法，研发高效能的智能化工具，对科技知识进行深入挖掘和重构，扩展实体、关系，促进知识体系的丰富化、细粒度化和语义化。

2.2　基本思路

基于上述的 3 个需求，笔者形成了科技大数据体系的 3 个建设思路。

2.2.1　扩展大数据资源体系，多维度丰富支撑数据群

基于原有的大数据体系，全面梳理并扩展权威、可获取的数据源，重新梳理基础数据将其扩展为以下 5 种：

（1）科研主体，包括专家学者、科研机构、学术期刊、科研团队、出版平台、科技企业与资助机构；

（2）科研活动，包括科研项目、学术会议、培训交流、科技大赛、数据分享、新闻资讯、社交活动与科技政策；

（3）科研成果，包括论文、专利、报告、获奖、专著、标准、软件、产品与数据；

（4）科研装置，包括大科学装置、仪器设备、耗材制剂、研究方法等；

（5）科学数据，包括研究数据等。

最终建立一个覆盖多类型、多渠道、多用户的包括文献、资讯、专业数据集、科研实体在内的完整的科技大数据生态体系。

2.2.2　以数据治理为核心，构建高效数据治理平台

改造原有的数据处理流程，以数据治理为核心，基于 AI 及大数据技术建设高效数据治理平台，实现科技大数据生态体系中的数据资源采集、数据存储、数据计算与数据管理的平台化运营，实现多来源数据组织和

融汇，形成科技文献大数据体系，有效提高各类资源的采集、汇聚、计算组织、融合与揭示能力。

2.2.3 基于智能化数据治理工具，建设高质量基础数据

面向智慧知识服务完成相关标准规范的制定和各种规范库的建设，为数据清洗、加工、组织提供基础保障；引入人工智能及语义技术，构建多样化的数据治理工具，实现多源异构数据汇聚、数据融汇、数据清洗、数据规范、知识抽取、关系抽取等，实现知识资源的有效治理，确保提供高质量数据资源。

2.3 以数据治理为核心的科技文献大数据体系框架设计

按照上面思路，笔者完成了科技文献大数据体系框架设计，如图 2 所示：

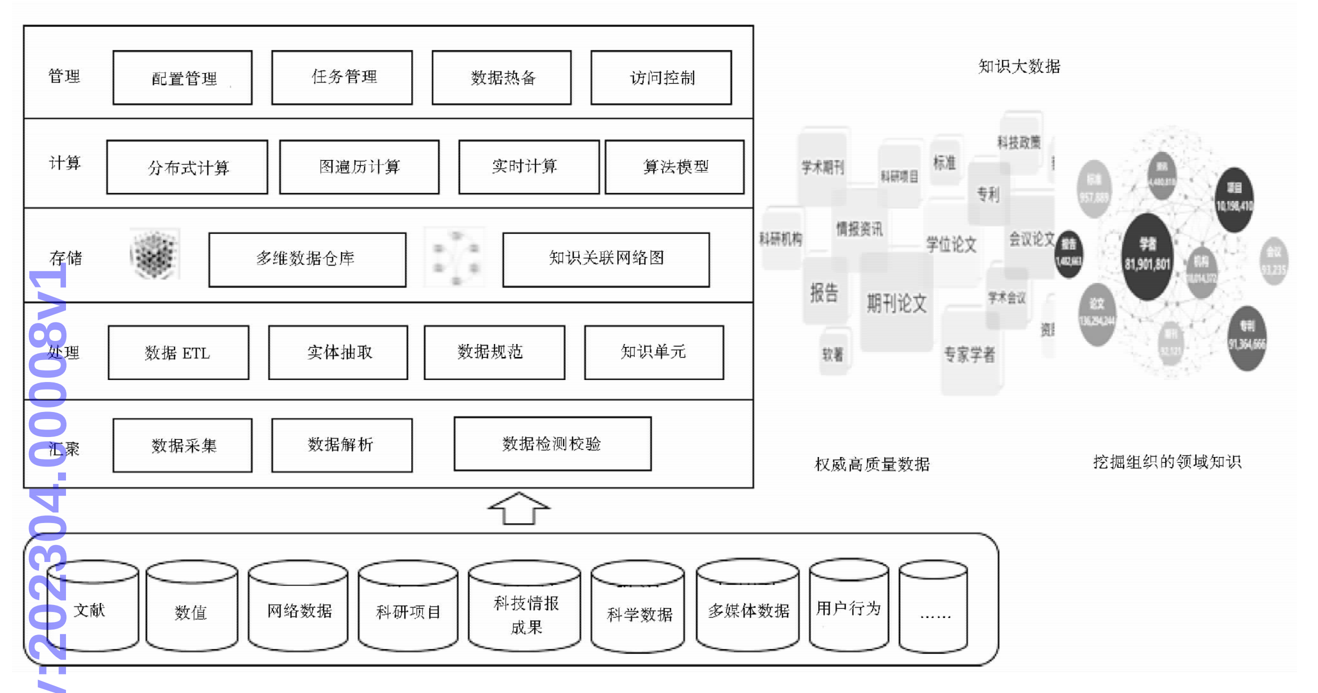


图 2 科技文献大数据体系框架设计

新的体系框架由 3 个部分组成：包括底层的基础文献数据、中间的数据治理平台以及经过治理后的应用数据层，包括高质量权威数据和领域知识数据。

从支撑智慧知识服务的角度，笔者根据数据的不同功能将数据体系划分 3 层，如图 3 所示：



图 3 科技文献大数据体系架构及建设成效

2.3.1 科技文献基础数据,即原始数据层

底层数据以文献及网络数据资源为主,是科技文献大数据体系最基础、最原始的数据资源,同时也是智能挖掘和分析的基础数据。

(1)商业出版资源:数据类型主要包括期刊论文、会议论文、学位论文、专利、科技情报、科技报告、标准、图书、期刊、工具书、产品样本、数值型数据集等。

(2)开放获取资源:通过官方的 OA 接口获取的期刊论文,主要包括 Cornell University Library 的 arXiv[10] 和美国国家生物技术信息中心的 PMC[11]。

(3)中国科学院(以下称简"中科院")体系资源:主要包括多年累积的近百家研究所的机构知识库信息,十三五规划中在建的数十个特色数据中心及专业中心所收集整理的大量多类型数据信息,其中包含文献情报体系自加工数据,这部分自加工数据多为 NSLC 各研究团队自己收集、加工、融汇的数据资源,资源类型覆盖较广,专业性领域性比较强。

(4)网络采集资源:主要为 NSTL 重点领域信息门户基于不同领域国内外相关机构网站,自动搜集、遴选、描述、组织和揭示各机构发布的重大新闻、研究报告、预算、资助信息、科研活动等内容。

(5)相关机构交换资源:主要包括期刊论文、会议论文、专利、科技情报信息、标准规范等。

2.3.2 数据治理基础数据

即数据治理层,主要包括用于数据质量控制的知识库数据。

(1)规范库。作为进行数据质量控制的传统方法,规范库依旧是科技文献大数据体系中的一个重要基础数据。大数据中心需要将分散在各中心、各团队、各项目中的规范库汇聚起来,通过统一管理集中服务来进一步推进协作共享,发挥价值。这些规范数据将被应用在数据清洗、加工、组织过程中,用于提升数据质量。主要包括:机构规范库、人名库、期刊库、基金项目库。

(2)领域词表。采用前期国家科技图书文献中心的 STKOS 项目[12] 以及 NSLC 的多年积累,汇聚覆盖理工农医四大领域的海量领域词表。

(3)知识库。知识库中主要包括三类信息。

其中多来源机构信息主要汇集了来自多个数据源的机构信息,包含中国科学院、国内主要研究机构及高校等 600 多家机构和研究所的基础机构信息,以及机构 IP 信息,机构订购商业出版社的信息。

多来源用户信息主要利用 WOS、iAuthor[13]、IR[14]、百度学术等多个来源的用户数据,以及中科院统一认证系统,配合相关信息服务系统所累积的用户信息,汇集整理用户基础信息库。

多平台日志信息包括研究所用户在使用商业出版商平台所反馈的多年累积的日志信息,以及 NSLC 自有各个服务平台的实时用户使用日志信息,以及对日志信息进行抽取、统计的后期数据信息。

(4)规则库。针对具体资源和资源特定属性的质量控制需求,建立特定的清洗规则组,用于支持智能化数据治理工具。

2.3.3 科技知识关联计算数据群

科技知识关联计算数据群即知识图谱层,是利用数据治理层的基础数据,将原始数据经过一系列的清洗、规范、融汇、抽取等处理而形成的不同类型的数据集合,面向应用服务层提供数据服务。其中实体数据包括从论文、项目、期刊、专利等数据资源中抽取的学者、机构、期刊、科研主题、会议、基金项目等科研实体。关系数据包括科研实体抽取时同时抽取的多种关系数据,形成实体关系库,在此基础上为关系挖掘与知识计算提供数据服务。

3 关键问题解决方案

基于上述设计方案,科技大数据体系还需要提供标注数据集、可靠的数据治理流程、多样化治理工具,同时要解决数据融汇、标引、领域知识图谱构建等难题。

3.1 覆盖数据生态全生命周期的精细化数据治理流程

3.1.1 覆盖数据生态全生命周期的治理流程

根据数据生态全生命周期的管理要求,笔者重塑了覆盖数据生态全生命周期的精细化数据治理流程,形成了包括数据源登记、数据收割、数据仓库、集成融汇、知识图谱、微服务 6 个主要阶段的标准化过程(见图 4),嵌入了基于大数据、机器学习、知识挖掘等技术开发的多种智能化治理功能模块,实现数据的精细化治理。

(1)数据源登记是落实数据源的甄选、接入方式、商务合作形式等基础信息。

(2)数据收割是根据数据源的释放方式进行对应的获取处理。目前主要的获取方式包括:OAI 接口访问、数据库直连访问、FTP 文件服务、存储介质手动获取。每类数据研发了匹配的配置模板,各数据源配置好目标字段在各个来源中的路径,便能进行新数据源的抽取,大大提升了接收效率。

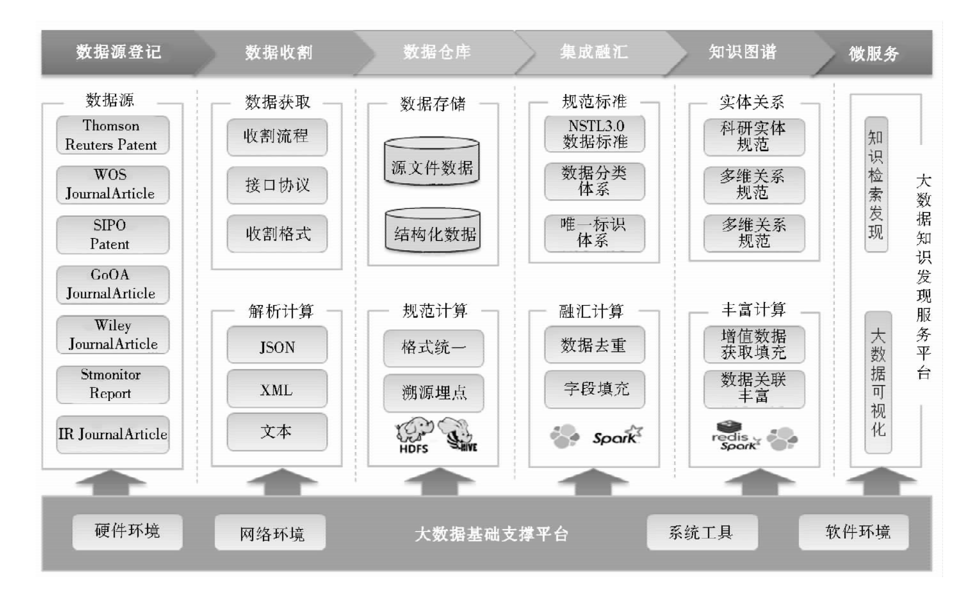


图4　高度自动化的数据汇聚与治理流程

(3)数据仓库是基于 Hive 数据仓库的外表式存储。通过分布式 MapReduce 进行 ETL 并行计算后的结构化数据存储在数据仓库中。数据存储是后续计算的数据基础。

(4)集成融汇是对解析后的结构化数据进行业务层面的去重和字符互补,并进行必要的信息转换和填充。目前论文的融汇规则是以标题＋期刊＋年份作为唯一识别法;专利则采用专利号、申请号;采集数据则采用 URL 的 md5 码。资源汇聚模块采用多级规则模式进行处理,面向大批量数据融汇更能体现其高效性。

(5)知识图谱构建包括数据丰富化、实体抽取、关系建立 3 个子过程。数据丰富化是为实体识别具有更好的精度,主要通过爬虫采集、定向加工、多源择优对比等手段完成数据丰富化处理。实体抽取主要是基于源数据进行处理,按照实体定义,通过抽取与分裂,构建实体对象。目前通过抽取得到的实体有学者、机构、期刊、科研主题、会议、基金项目六大类。关系构建是知识图谱的重要环节,在完成文献中实体分离后,保留实体之间的关系,并通过对关系数据的统计分析完成权重计算,固化实体间关系的权重值。

(6)微服务。科技文献大数据体系通过 Restful API 接口提供数据获取服务。目前采用分布式技术,具有弹性扩展性、热注册、高性能、防爬虫等优点。

3.1.2　可个性化配置的模块化流程

笔者将每个处理过程进行模块化设计,可以面向多样化数据来源和格式构建不同的收割、解析、清洗等功能模块,使得每种数据资源的处理过程均可实现个性化配置。

同时对不同的资源类型提供了不同的 package 标识,利用不同类去实现不同的来源收割。每新增一种数据源,按照类型收割的配置定义规则,对收割频率、收割字段、收割类型(如增量、全量)、收割开始时间等进行配置,就能实现新数据源收割。

3.1.3　可视化的全自动处理流程

在实现上,采用 MapReduce 和 Spark 框架实现分布式计算处理。每一个完整处理流程都可配置为一个作业工程,通过设定作业队列临界值,将数据处理作业分摊给多个服务器同步处理,实时动态加载。平台还通过可视化方式展示处理流程中的各个步骤,如"未处理""处理中""已处理"等。同时对某一来源数据可以进行全量和次增量的重跑。这种自动化流程处理降低了工作复杂程度,保证了一定程度的个性化,还提高了安全性。

通过上述数据处理建立起从数据资源接收登记、存储管理、审计校验、运行监控、使用管理、备份管理的全生命周期的标准管理流程,并形成一系列管理规范,以此来约束和保障各类科学数据在接收、校验、存储、使用、备份的正常状态和使用规范。

3.2　多重数据质量控制

智慧知识服务的主要特征即是个性化和精准化,这两项都需要高质量数据支持。笔者通过分析数据特征、进行数据标准化、对数据质量进行监控和校验,实施多重数据质量控制,切实改善数据质量和可靠性。

3.2.1 统一的元数据标准和元数据模型

科技文献大数据体系基于 NSTL 统一文献元数据标准 3.0(正式版)[15]制定了大数据体系统元数据标准,采用 XML 语言和 DTD 分别对标准进行了形式化描述,元数据共包含 13 个元素集:来源、单篇文献、主题/分类/关键词、贡献者/机构、会议、基金、操作信息、获取管理、全文文件、图、表、附加资料和参考文献元素集。不计重复元素和属性,本标准共包含 97 个描述性元素、53 个辅助性元素、49 个属性以及 4 个特殊字符元素。通过元素和属性的灵活组合来描述多样化、多层次的资源。

3.2.2 规范库建设

通过整合汇集来自不同机构和项目的规范库数据,目前形成包括机构规范库、人名库、期刊库、基金项目库共 4 种实体规范库。在数据清洗过程中,用于规范相关元数据内容。

3.2.3 多重数据清洗规则

作为规范库的补充,还增加了针对具体资源和资源特定属性的质量控制。每种类型资源建立特定的清洗规则组,同时还依据相关的标准建立了面向特定元素的清洗规则,对来源国家/地区、城市、机构名称、期刊名称、出版年、数据类型字段、学科分类信息、学者姓名、关键词等字段进行规范。

3.2.4 可重复的清洗过程

科技文献大数据体系基于上述标准规范和规范库,通过标准 API 接口对数据采集、汇聚、清洗、加工、组织、计算、服务等完整数据质量生命周期进行监控管理。同时平台提供了可重用的清洗流程,保障数据各个阶段的重复可操作,以循环提升数据质量。

3.2.5 融入专家智慧的数据加工工具与工作机制

科技文献大数据体系以规范库和第三方资源作为计算的基础依据,对原始数据进行一系列自动化处理,同时采用融入专家智慧的数据加工工具。建立了主要通过数据管理加工系统实现对机构、学者、期刊、主题词等进行深加工,同时由专业人员参与以确保数据质量有效控制的工作机制。

3.3 面向多源科技大数据的数据融合和数据标引

3.3.1 数据融合

数据融合是集机器自动融合和人工治理融合相结合的数据治理过程,集 ETL 流程化、实体规范化、数据去重和丰富化的数据治理标准化流程为一体,即是一种基于规则算法的数据定制化融合流程,也是一种群智群策的数据融合流程。

每种实体类型的数据都有各自的排重要素和排重规则,以期刊论文为例,首先以文献 DOI 为第一排重要素,其次以文献标题 + 文献作者数目 + 文献作者姓名 + 文献出版年份的组合为第二排重要素,进一步完成数据排重,为数据融合打好基础。然后,基于大数据平台的 MapReduce,Spark 等高性能计算技术为计算引擎,Hive 等数据仓库为数据源,Elasticsearch 等高速度的服务索引为依托,完成数据识别与融合过程,如图 5 所示:

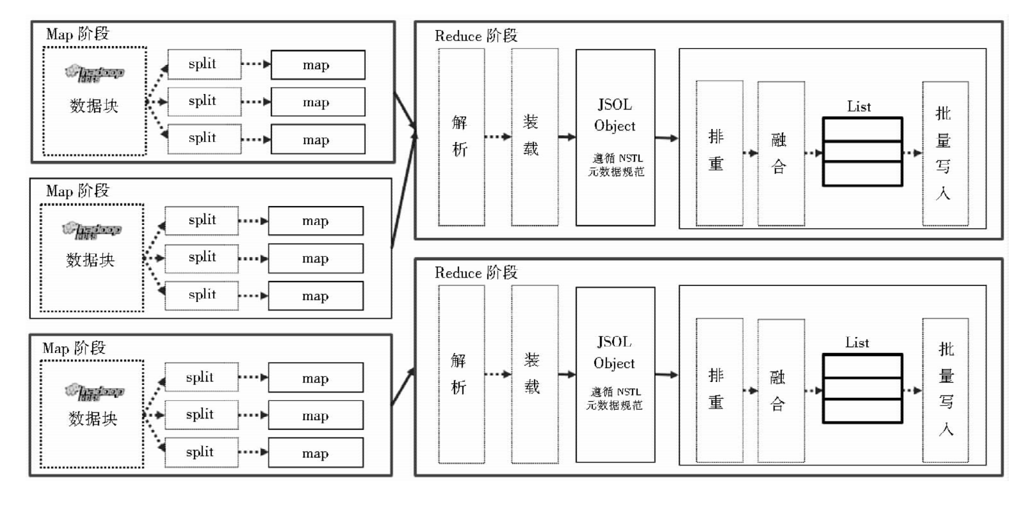


图 5 基于 Mapreduce 的数据识别与融合流程

3.3.2 数据标引

数据集的分类与标引是科学研究过程中的智慧化体现,也是数据公共服务的关注点。对高价值的科技论文、专利、期刊、科技报道、研究报告等多类型知识资源,科技文献大数据体系实现了从主题关键词、学科分类、重要度、发表时效等多个维度的标引。以学科分类标引为例,科技文献大数据体系研究了多个来源学科分类体系,如 NSTL、中图、科图、ESI 以及出版商学科分

类,对多级别学科进行综合分析与计算融合,形成适于科技文献大数据体系的学科分类库。然后,将此学科分类库应用于体系内的各类资源,每条数据资源都"烙印"自己的学科分类集。最后,以学科分类为核心,提供对外服务和使用。

科技文献大数据体系基于人工智能及语义技术开发了多样化数据治理工具,为各类实体制定了个性化的标引规则,实现了计算机自动化标引,同时也开发了融入专家智慧的数据加工平台,建立了协调参与的数据质量控制的工作机制,使得学者、机构管理者、学科服务团队、数据管理团队等不同角色用户能够参与数据治理,实现智能、精准的专题领域画像,为精准推送、精准检索等精准服务提供高质量高价值的数据标引知识库。

4 建设成效及应用介绍

经过3年多的建设,目前已形成了多领域与多层次的科技文献大数据知识资源体系[16]。相比传统的文献大数据体系,其所包括的数据内容更为丰富,除了传统文献数据资源,拥有了更为丰富的数据治理数据以及科技知识关联计算数据,形成了覆盖数据生态全生命周期的精细化数据治理流程,嵌入多个智能化治理模块及工具,能够为智慧知识服务提供高质量的知识数据。

4.1 初步形成一定体量的多个数据群

4.1.1 科技文献基础数据群

目前该类数据已经覆盖了Web of Science、Elsevier、Willey、Taylor、维普、CSCD、PMC、arXiv等8家国内外知名数据库;同时采集了来自NIH、NSF、NSFC等国内外2 200余个重要科研机构的数据,数据总量已超3亿。

同时该类数据除了传统的期刊论文、图书、专利、学位论文、科技报告、标准以及古籍等类型的数据资源,还覆盖了全球基金项目、全球重要科研机构以及学协会的全网络科技数据、全球重要科技智库的开放科技网络数据、社会经济信息数据、政策法规信息数据、来自世界银行以及洛桑报告的科技竞争力的数值型数据、收集汇聚了中科院重要知识服务系统的用户行为数据等。

总的来看,该数据群年度跨度大,鲜活度高。该体系的数据最早回溯到1799年(专利)和1900年(文献),数据的时间跨度长达221年。数据定期更新频率为1天(文献)和3天(专利),以确保数据鲜活度。

4.1.2 数据治理基础数据

目前已经累积规范库和领域词表两大类数据,如表1所示:

表1 数据治理基础数据群的支撑数据资源

基础数据群	分类	明细
规范库	机构规范库	CSCD、iSwitch、patent等来源机构规范名称约9万条,机构别称约90万条
		中国科学院机构网站和中国科学院机构知识库的中国科学院机构
		中华人民共和国教育部网站的中国高等院校[17]
		全球研究标识符(Grid)网站的全球221个国家近8万个机构
		维基百科DBPedia的全球大学数据[18]
	学者规范库	WOS、iAuthor、IR、百度学术等来源的学者约186万条,关联文献资源420万篇
		中国科学院各研究所官网
		中国科学院机构知识库
		中国科学家在线[19]
	期刊规范库	全国联合期刊目录知识库和自动采集的期刊数据,约4.5万篇
	基金项目规范库	澳大利亚,德国,俄罗斯,加拿大,美国,欧盟,日本,瑞士,印度、英国,中国等共计11个国家的基金项目,约500万个
领域词表	理	344 735个术语,104 063个概念
	工	605 604个术语,157 570个概念
	农	241 530个术语,92 869个概念
	医	1 128 835个术语,260 329个概念

4.1.3　科技知识关联计算数据群

该类数据目前已累积了包括学者、机构、论文、专利、期刊、基金项目科研实体超过 3 亿个。同时还累积六类科研实体的 21 类关系数据 34 亿对，形成了一个相对全面的实体关系库，可以为关系挖掘与知识计算提供智慧数据服务，能够构建面向学术研究圈的权威知识图谱，并为智慧知识服务提供强力知识基础，如表 2 所示：

表 2　21 类科研实体关系数据

序号	英文名	约束 - 关系类型	主体	客体
1	publish	出版关系	机构	文献集
2	address_is	地址关系	研究者/机构/会议	国家/州省/城市
3	source_is	来源关系	单篇文献	文献集
4	subject_is	主题分类关系	文献集/单篇文献/机构/研究人员/项目	主题/分类/关键词
5	contributor	贡献关系	文献集/单篇文献	研究者
6	affiliation	所属关系	研究者	机构
7	proceeding_include	会议收录关系	会议	文献集/单篇文献
8	hold_coference	举办关系	机构	会议
9	fund_by	资助关系	项目	机构
10	reference	引用关系	文献集/单篇文献	文献集/单篇文献
11	hold_collection	收藏关系	文献集/单篇文献	数据库
12	attach_with	附件关系	文献集/单篇文献/研究者/机构	全文/图/表/附加材料
13	fundapply	申请关系	项目	研究者
14	manageby	上级机构	子机构	机构
15	related_org	相关关系	机构	机构
16	contribute_institution	贡献机构关系	文献集/单篇文献	研究机构
17	undertake_conference	承办关系	机构	会议
18	support_conference	支持关系	机构	会议
19	cooperate_conference	协办关系	机构	会议
20	guid_conference	指导关系	机构	会议
21	associatemedia_conference	合作媒体	机构	会议

4.2　初步形成支撑智慧知识服务能力

目前科技文献大数据体系建设初见成效，已经为 NSLC 门户网站[20]以及“慧”系列产品等多个服务[21-23]提供数据支持，同时提供多种形式的数据服务[24]。

以机构知识管理与数据分析服务[25]为例，科技文献大数据体系为该服务提供了科技文献基础数据群和知识关联计算数据群，以支撑服务按照机构维度进行自动汇聚科研机构科技成果数据、智能计算与描绘机构学术画像以及机构当前布局情况及发展方向。同时还支撑服务实时提供该机构的研究人员数据、科研基金项目数据、发表期刊论文数据等，见图 6。

5　结语

随着科技文献大数据体系为越来越多应用服务提供数据支撑，一些潜在的问题也逐渐显露出来，需要笔者在今后的建设过程中进行认真的思考和解决。

首先是可持续发展的问题，需要笔者认真分析所涉及的所有资源渠道保障，分析各类型资源的可能获得来源、可能的保障方法和机制，寻找适当的运作模式，在有限资金的投入下，以共建共享、数据服务等多种机制，激励多来源数据的提供者的参与贡献。同时面向不断变化的应用需求，采用大数据和 AI 技术，基于原有数据源进行深度挖掘、发现，实现数据资源的增值，促进大数据体系的不断丰富化。

其次需要持续提升科技文献大数据体系的质量控制能力。由于数据来源多，数据质量不一、遵循的标准不一，在数据清洗和融汇的过程中存在很多隐藏的问题，影响了数据融汇准确性和数据组织的规范性，也影响了对于顶层智能知识服务的各种应用的效果。后续笔者还需要加强对规范库的维度、层次的丰富化，结合新技术新模型的应用，有效提升数据的完整性和质量。

吴振新，钱力，谢靖，等. 面向智慧知识服务的科技文献大数据体系建设[J]. 图书情报工作,2020,64(24):63 – 72.

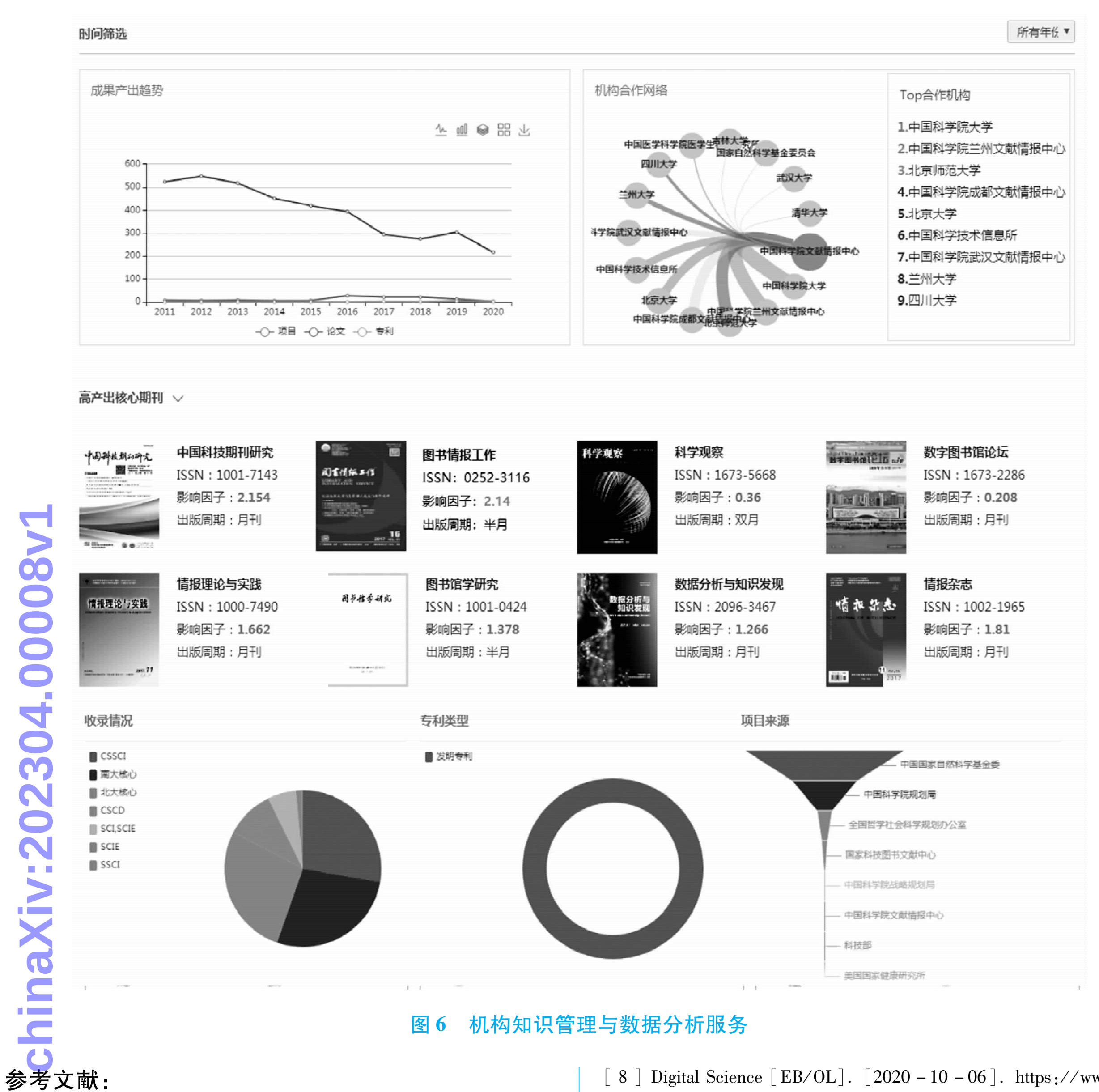


图 6　机构知识管理与数据分析服务

参考文献：

[1] 人工智能那么火～如今 AI 的应用场景都有哪些? [EB/OL]. [2020 – 11 – 16]. https://www.zhihu.com/question/282715644.

[2] The AI Hierarchy of Needs [EB/OL]. [2020 – 11 – 16]. https://hackernoon.com/the-ai-hierarchy-of-needs-18f111fcc007.

[3] AIBigbull2050. 人工智能三驾马车:算法,算力,算据 [EB/OL]. [2020 – 11 – 16]. http://blog.itpub.net/69946223/viewspace-2734390.

[4] 刘琳. 新基建热潮下,AI 基础数据服务会有哪些变化? [EB/OL]. [2020 – 06 – 03]. https://www.leiphone.com/news/202006/WNW3OH7baaG0RBi5.html.

[5] 钱力，张晓林，王茜. 基于科技文献的研究设计指纹描述框架研究[J]. 大学图书馆学报，2015，33(1)：14 – 20.

[6] 柯平,邹金汇. 后知识服务时代的图书馆转型[J]. 中国图书馆学报，2019,45(1):4 – 17.

[7] Research Intelligence [EB/OL]. [2020 – 10 – 06]. https://www.elsevier.com/research-intelligence.

[8] Digital Science [EB/OL]. [2020 – 10 – 06]. https://www.digital-science.com.

[9] Wizdom.ai[EB/OL]. [2020 – 10 – 06]. https://www.wizdom.ai/#about.

[10] arXiv[EB/OL]. [2020 – 10 – 06]. https://arxiv.org.

[11] Pubmed[EB/OL]. [2020 – 10 – 06]. https://www.ncbi.nlm.nih.gov/pubmed.

[12] STKOS 科技知识组织体系共享服务系统[EB/OL]. [2020 – 10 – 06]. http://stkos.las.ac.cn/stkosservice/user/welcome.htm.

[13] iAuthor 中国科学家在线[EB/OL]. [2020 – 10 – 08]. http://iauthor.cn/welcome/index.

[14] 中国科学院机构知识库网格[EB/OL]. [2020 – 10 – 08]. http://www.irgrid.ac.cn.

[15] NSTL 统一文献元数据标准 3.0 [EB/OL]. [2020 – 10 – 08]. http://spec.nstl.gov.cn/embed/home.htm.

[16] SciFire 基于群体智能的知识服务平台[EB/OL]. [2020 – 10 – 08]. http://159.226.100.96/bi/bi.html.

[17] 中华人民共和国教育部网站提供的中国高等院校的名单[EB/OL]. [2020-10-08]. http://www.moe.gov.cn/srcsite/A03/moe_634/201706/t20170614_306900.html.
[18] DBPedia 的全球大学数据[EB/OL]. [2020-10-08]. https://wiki.dbpedia.org/develop/datasets.
[19] 中国科学家在线[EB/OL]. [2020-10-08]. https://iauthor.cn.
[20] 中国科学院文献情报中心中国科学院知识服务平台[EB/OL]. [2020-10-08]. https://www.las.ac.cn.
[21] 科技大数据知识发现平台[EB/OL]. [2020-10-10]. https://scholareye.cn/.
[22] 慧科研个人版智能随身科研助理[EB/OL]. [2020-10-10]. https://scholarin.cn/.
[23] 中国科学院文献情报中心数据观测平台[EB/OL]. [2020-10-10]. http://kgview.las.ac.cn.
[24] 中国科学院文献情报中心中国科学院知识服务平台数据服务[EB/OL]. [2020-10-10]. https://www.las.ac.cn/front/dataCenter/dataResources.
[25] 慧科研机构版机构知识管理与分析服务平台[EB/OL]. [2020-10-10]. https://inst.scholarin.cn/.

作者贡献说明：
吴振新：负责科技文献大数据体系设计与建设，论文撰写与最终版本修订；
钱力：参与科技文献大数据体系设计，负责大数据基础技术平台建设；
谢靖：参与科技文献大数据体系设计，负责科技大数据管理平台建设；
常志军：参与科技文献大数据体系设计，负责大数据基础技术平台实施；
许丽媛：汇聚、清洗数据，参与制定数据清洗规则，论文初稿撰写及修改；
赵艳：参与科技文献大数据体系设计。

Construction of Sci-Tech Big Data System oriented to Intelligent Knowledge Service

Wu Zhenxin[1,2] Qian Li[1,2] Xie Jing[1,2] Chang Zhijun[1,2] Xu Liyuan[1] Zhao Yan[1,2]
[1] National Science Library, Chinese Academy Sciences, Beijing 100190
[2] Department of Library, Information and Archives Management, School of Economics and Management, University of Chinese Academy of Sciences, Beijing 100190

Abstract: [**Purpose/significance**] The paper explores the construction of literature intelligence big data knowledge resource system, which supports multi-domain intelligent knowledge service. [**Method/process**] Based on the AI application requirements, drawing on the industry experience, combing the problems of existing resource system, the paper expanded the resource system from multi-level and multi-dimensional, built a reliable data processing process and computing platform to support efficient data collection and processing, and developed intelligent data governance tools to achieve effective governance of knowledge resources and ensure the provision of high-quality data resources. [**Result/conclusion**] It has initially formed a knowledge resource system covering multiple types and disciplines of sci-tech literature, constructed and completed a highly automated data collection and governance process, implemented multiple data quality control, and accumulated hundreds of millions of high-quality data. At present, it has provided data support for multiple knowledge services.

Keywords: science and technological big data knowledge resource architecture data aggregation intelligent knowledge services